International Journal of Population Data Science

Journal Website: www.ijpds.org

Bringing more science to the art of linkage: Using positive predictive value of weight and outcome sets to reduce subjectivity in probabilistic linkage

Hills, Brent[1]

[1]Population Data BC

Objectives

Currently, a probabilistic linkage is performed by our organization with final linkage classification established by users with expert knowledge applying rules referencing weight and comparison outcome sets. The particular classification results in a perceived comprehensive linkage. Acknowledged weaknesses are variation in expert knowledge and its application. Also, consideration of expert rules is often time-consuming.

We piloted a new approach, involving a file level summary of "positive predictive value" for weights and outcome sets. We contrast the new approach with the previous one and identify strengths and weaknesses.

Approach

We resolve linkages using two different approaches, the existing method that expert users apply rules to weight and comparison outcomes, and a recent one using positive predictive values (PPV) that reduce resolution subjectivity.

The new method produces summary true positive, false positive and positive predictive values for each weight and outcome set within a file of candidate pairs above the cutoff. The top-weighted pair increments the true positive (TP) for the weight and outcome set on the record. All other candidate pairs increment false positive counts (FP). At the end of the file, PPV is calculated for each weight and outcome as TP/(TP+FP). Additionally, a .9 PPV weight threshold is established from the summary excluding weights with less than N occurrences. The approach presumes successful link

Accepted links include top-ranked records whose weight is greater than the .9 PPV threshold and top candidates with an outcome summary PPV >= .9, excluding outcome sets with less than N occurrences.

Results

In the pilot, the new method produced linkage results near par with the previously employed methods. Importantly, they establish accepted links by a consistent methodology, allowing for increased standardization. The method eases identification of a threshold weight and referencing summary comparison outcome PPVs identifies additional confident links in larger population-based linkages which a single weight threshold may exclude. Components require minimal tuning to data characteristics, error tolerance, and result expectations. The new method is unable to identify links established in the legacy method by additional processing or manual review. Further review of the approach as applied to other datasets is needed.

*Corresponding Author:
Email Address: brent.hills@popdata.bc.ca (B. Hills)

http://dx.doi.org/10.23889/ijpds.v1i1.193
August 2016 © The Authors. Open Access under CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en)